Dinucleotides and the arrow of time

or:

An irreversible context-dependent substitution model

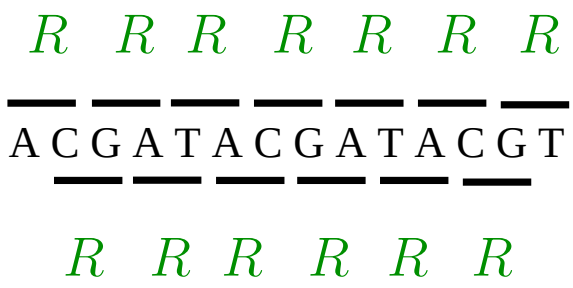
Gerton Lunter
MRC Functional Genetics Unit
University of Oxford

Model for context-dependent mutations

- Model: continuous-time finite-state Markov chain
- Parameters: 16x16 rate matrix R, for all dinuc-to-dinuc substitutions.
- Process: Let R act on all (overlapping) dinucleotides simultaneously.

Model for context-dependent mutations

- Model: continuous-time finite-state Markov chain
- Parameters: 16x16 rate matrix R, for all dinuc-to-dinuc substitutions.
- Process: Let R act on all (overlapping) dinucleotides simultaneously.



Problem: contagious dependence

For sequence of length L, state space has 4^L states. Matrix has size $4^L \times 4^L$.

Model for dependent mutations

Full model has state space of size 4^L.

Let D_i be matrix acting on full state space, but with R acting on single pair of sites $i, i+1$.

Rate matrix S acting on entire sequence is

$$S = \sum_{i=1}^{L-1} D_i$$

Model for dependent mutations

To calculate probabilities:

$$\exp(St) \quad = \quad \exp\left(t\sum_{i=1}^{L-1} D_i\right) \quad = \quad \sum_{n=0}^{\infty} \frac{t^n}{n!}\left(\sum_{i=1}^{L-1} D_i\right)^n$$

Many terms in product $\left(\sum_{i=1}^{L-1} D_i\right)^n$ commute.

Strategy:
Keep commuting terms, throw away high order non-commuting terms.

Result:
Dynamic programming algorithm approximating $\exp(St)_{p,q}$.

Likelihood algorithm - some notation

p, q, s, t Sequences

Likelihood algorithm - some notation

p, q, s, t	Sequences
ps	Concatenation of p and s

Likelihood algorithm - some notation

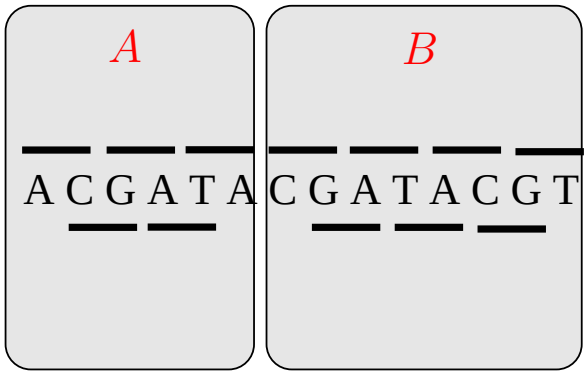
p, q, s, t	Sequences
ps	Concatenation of p and s
$A_{p,q}$	Matrix element

Likelihood algorithm - some notation

p, q, s, t	Sequences
ps	Concatenation of p and s
$A_{p,q}$	Matrix element
$A \oplus B$	Matrix concatenation sum

$$(A \oplus B)_{ps,qt} = A_{p,q}\delta_{s,t} + \delta_{p,q}B_{s,t}$$

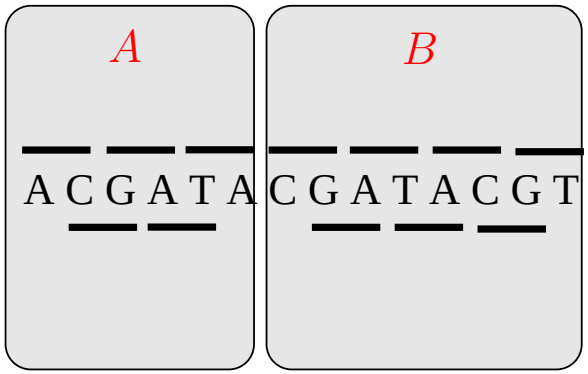
$A \oplus B$:



Likelihood algorithm - some notation

p, q, s, t	Sequences	
ps	Concatenation of p and s	
$A_{p,q}$	Matrix element	
$A \oplus B$	Matrix concatenation sum	$(A \oplus B)_{ps,qt} = A_{p,q}\delta_{s,t} + \delta_{p,q}B_{s,t}$
$A \otimes B$	Matrix concatenation product	$(A \otimes B)_{ps,qt} = A_{p,q}B_{s,t}$

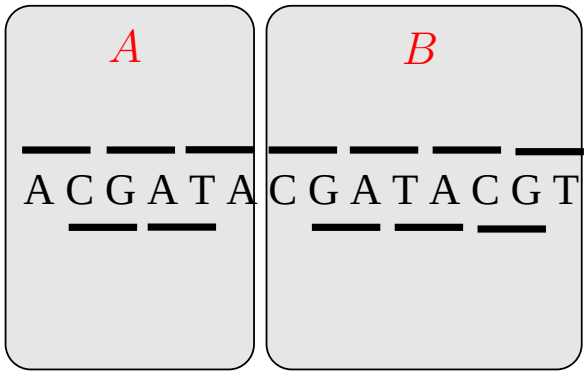
$A \oplus B$:



Likelihood algorithm - some notation

p, q, s, t	Sequences	
ps	Concatenation of p and s	
$A_{p,q}$	Matrix element	
$A \oplus B$	Matrix concatenation sum	$(A \oplus B)_{ps,qt} = A_{p,q}\delta_{s,t} + \delta_{p,q}B_{s,t}$
$A \otimes B$	Matrix concatenation product	$(A \otimes B)_{ps,qt} = A_{p,q}B_{s,t}$
		$\exp(A \oplus B) = \exp(A) \otimes \exp(B)$

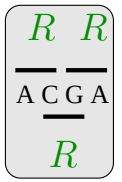
$A \oplus B$:



Divide and conquer

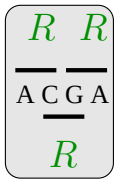
definition: $S_k = \sum_{i=1}^{k-1} D_i$ $(4^k \times 4^k$ matrix)

e.g. S_4 :



Divide and conquer

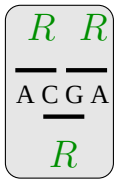
definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 :



$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

$$D_i D_j = D_j D_i \text{ unless } |i-j| = 1$$

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

$$D_i D_j = D_j D_i \text{ unless } |i-j| = 1$$

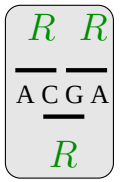
Def.: $D_{i_1} D_{i_2} \cdots D_{i_n}$ is overlapping if not the product of commuting factors.
The length of an overlapping term is the number of sites it affects.

Examples:

$D_1 D_3 D_2$ overlapping, length 4.

$D_1 D_4 D_2 D_5 = (D_1 D_2)(D_4 D_5)$ not overlapping.

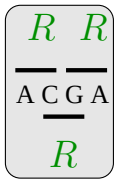
Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right) \frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

Idea: Factorize terms of expansion into commuting overlapping factors.

Divide and conquer

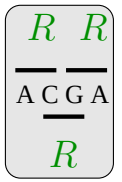
definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right) \frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

Idea: Factorize terms of expansion into commuting overlapping factors.

Consider overlapping factor F of length k containing 'rightmost' D_{L-1}.

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

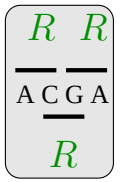
$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right) \frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

Idea: Factorize terms of expansion into commuting overlapping factors.

Consider overlapping factor F of length k containing 'rightmost' D_{L-1}.

- F contains only factors D_i with $i > L - k$ (and contains all of those).

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

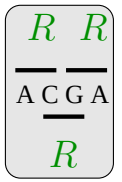
$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

Idea: Factorize terms of expansion into commuting overlapping factors.

Consider overlapping factor F of length k containing 'rightmost' D_{L-1}.

- F contains only factors D_i with $i > L - k$ (and contains all of those).
- All terms containing F in expansion can be written GF.

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right) \frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

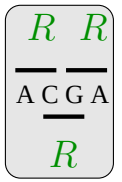
Idea: Factorize terms of expansion into commuting overlapping factors.

Consider overlapping factor F of length k containing 'rightmost' D_{L-1}.

- F contains only factors D_i with $i > L - k$ (and contains all of those).
- All terms containing F in expansion can be written GF.
- This product commutes, since G only contains D_i with $i < L - k$.

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 :



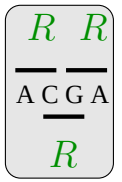
$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

Idea: Factorize terms of expansion into commuting overlapping factors.

Consider overlapping factor F of length k containing 'rightmost' D_{L-1}.

- F contains only factors D_i with $i > L - k$ (and contains all of those).
- All terms containing F in expansion can be written GF.
- This product commutes, since G only contains D_i with $i < L - k$.
- Suppose F has degree n. Can we compute G?

Divide and conquer

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\exp(S_L t) = I_L + \left(\sum_{i=1}^{L-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-1} D_i\right)^2 \frac{t^2}{2!} + \cdots$$

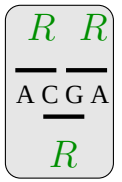
$$G = I_{L-k}\frac{t^n}{n!} + \left(\sum_{i=1}^{L-k-1} D_i\right)\frac{\binom{n+1}{1} t^{n+1}}{(n+1)!} + \left(\sum_{i=1}^{L-k-1} D_i\right)^2 \frac{\binom{n+2}{2} t^{n+2}}{(n+2)!} + \cdots$$

$$= \frac{t^n}{n!}\left[I_{L-k} + \left(\sum_{i=1}^{L-k-1} D_i\right)\frac{t}{1!} + \left(\sum_{i=1}^{L-k-1} D_i\right)^2 \frac{t^2}{2!} + \cdots\right]$$

$$= (t^n/n!)\exp\left[(S_{L-k}t \oplus O_k)t\right]$$

$$= \exp(S_{L-k}t) \otimes (\frac{t^n}{n!} I_k)$$

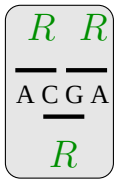
Likelihood algorithm

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

{Collect length-k F's in A_k} $\Rightarrow$

$$\begin{aligned} \exp(S_L) = {} & \exp(S_{L-1}) \otimes A_1 + \\ & \exp(S_{L-2}) \otimes A_2 + \\ & \vdots \\ & \exp(S_1) \otimes A_{L-1} + A_L \end{aligned}$$

Likelihood algorithm

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

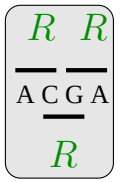
$$\{\text{Collect length-}k\text{ F's in } A_k\} \quad \Rightarrow \quad \begin{aligned} \exp(S_L) \;=\; & \exp(S_{L-1}) \otimes A_1 + \\ & \exp(S_{L-2}) \otimes A_2 + \\ & \vdots \\ & \exp(S_1) \otimes A_{L-1} \;+\; A_L \end{aligned}$$

- Recursively solve for A_k:

$$\begin{aligned} A_1 &= \exp(S_1) \\ A_2 &= \exp(S_2) - A_1 \otimes A_1 \\ A_3 &= \exp(S_3) - A_1 \otimes A_2 - A_2 \otimes A_1 - A_1 \otimes A_1 \otimes A_1 \\ &\vdots \end{aligned}$$

- Matrix A_k is of dimension $4^k \times 4^k$.
- Matrix A_k contains terms of degree $k-1$, and h.o.t.

Likelihood algorithm

definition: $S_k = \sum_{i=1}^{k-1} D_i$ ($4^k \times 4^k$ matrix) e.g. S_4 : 

$$\{\text{Collect length-}k\text{ F's in } A_k\} \quad \Rightarrow \quad \begin{aligned} \exp(S_L) \;=\; & \exp(S_{L-1}) \otimes A_1 + \\ & \exp(S_{L-2}) \otimes A_2 + \\ & \vdots \\ & \exp(S_1) \otimes A_{L-1} \;+\; A_L \end{aligned}$$

- Recursively solve for A_k:

$$\begin{aligned} A_1 &= \exp(S_1) \\ A_2 &= \exp(S_2) - A_1 \otimes A_1 \\ A_3 &= \exp(S_3) - A_1 \otimes A_2 - A_2 \otimes A_1 - A_1 \otimes A_1 \otimes A_1 \\ &\vdots \end{aligned}$$

- Matrix A_k is of dimension $4^k \times 4^k$.
- Matrix A_k contains terms of degree $k-1$, and h.o.t.

Dynamic programming: Recursively compute matrix elements

$\exp(S_1)_{p_1\ ,\ q_1}$, $\exp(S_2)_{p_1p_2\ ,\ q_1q_2}$, $\ldots$ $\exp(S_k)_{p_1p_2\cdots p_k\ ,\ q_1q_2\cdots q_k}$

Dinucleotide model

Full model: arbitrarily parameterized matrix R, 240 parameters.

We used a subset:

- Assume strand symmetry
  (e.g. same rates for CG $\rightarrow$ TG and CG $\rightarrow$ CA substitutions).
- One parameter for all dinucleotide mutation rates (e.g. AT$\rightarrow$GC).

In all $48 + 1$ parameters.

Parameter inference

Inference was done by Bayesian MCMC sampling.

Advantages of MCMC over ML:

- Easier to program
- More robust against local maxima
- Confidence intervals for free

Parameter inference

Inference was done by Bayesian MCMC sampling.

Advantages of MCMC over ML:

- Easier to program
- More robust against local maxima
- Confidence intervals for free

For this experiment:

- Input: Aligned sequences of about 100,000 bp
  (2 sets: nongenic human-mouse DNA and synthetic data)
- 600,000 iterations
- ESS in range 100-500 (depending on parameter)

Results

	*			
	A	C	G	T
A	*	.13	.12	.12
C	.11	*	.11	.14
G	.14	.11	*	.11
T	.12	.12	.13	*

	A			
A	C	G	T	
*	.03	.01	.00	
.01	*	.03	.02	
.02	.01	*	.02	
.03	.00	.01	*	

	C			
A	C	G	T	
*	.00	.01	.02	
.02	*	.04	.01	
.01	.01	*	.02	
.01	.01	.01	*	

	G			
A	C	G	T	
*	.01	.02	.02	
.03	*	.01	2.30	
.01	.01	*	.00	
.01	.02	.02	*	

	T			
A	C	G	T	
*	.01	.01	.02	
.03	*	.03	.01	
.01	.01	*	.02	
.01	.03	.01	*	

Total mononucleotide rate: 0.509 ± 0.007 (true 0.502); total dinucleotide rate: 0.018 ± 0.003 (true 0.020)

CG→TG rate: 2.47 ± 0.09 (true 2.40)

Relative deviation from 0: : <2.2 : 2.2–3.3 : 3.3–4.4 : 4.4–5.5 : 5.5–6.6 : >6.6 std. devs.

Results

	*			
	A	C	G	T
A	*	.13	.12	.12
C	.11	*	.11	.14
G	.14	.11	*	.11
T	.12	.12	.13	*

	A				C				G				T			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	*	.03	.01	.00	*	.00	.01	.02	*	.01	.02	.02	*	.01	.01	.02
C	.01	*	.03	.02	.02	*	.04	.01	.03	*	.01	2.30	.03	*	.03	.01
G	.02	.01	*	.02	.01	.01	*	.02	.01	.01	*	.00	.01	.01	*	.02
T	.03	.00	.01	*	.01	.01	.01	*	.01	.02	.02	*	.01	.03	.01	*

Total mononucleotide rate: 0.509 ± 0.007 (true 0.502); total dinucleotide rate: 0.018 ± 0.003 (true 0.020)

CG→TG rate: 2.47 ± 0.09 (true 2.40)

	*			
	A	C	G	T
A	*	.05	.11	.02
C	.02	*	.04	.15
G	.15	.04	*	.02
T	.02	.11	.05	*

	A				C				G				T			
	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
A	*	.00	.00	.00	*	.00	.13	.03	*	.03	.11	.06	*	.03	.11	.03
C	.04	*	.01	.06	.03	*	.00	.22	.10	*	.08	2.47	.00	*	.01	.00
G	.12	.03	*	.04	.04	.06	*	.05	.00	.00	*	.00	.16	.05	*	.06
T	.05	.09	.01	*	.00	.00	.00	*	.00	.06	.02	*	.02	.08	.00	*

Total mononucleotide rate: 0.469 ± 0.005; total dinucleotide rate: 0.016 ± 0.002

CG→TG rate: 2.55 ± 0.12

Relative deviation from 0: : <2.2 : 2.2–3.3 : 3.3–4.4 : 4.4–5.5 : 5.5–6.6 : >6.6 std. devs.

Model comparison

Model:	Params	Likelihood	Tot. mononuc.	$CG \rightarrow TG$	Tot. dinuc.
GenRev	7	-228397.0 $\pm$ 2.2	.395 $\pm$ .003	–	–
GenIrr	12	-228392.9 $\pm$ 2.5	.396 $\pm$ .003	–	–
GenIrr+CG	13	-224958.7 $\pm$ 2.2	.420 $\pm$ .003	1.50 $\pm$ 0.02	–
GenIrr+CG+Dinuc	14	-224751.7 $\pm$ 3.1	.368 $\pm$ .003	1.43 $\pm$ 0.03	.025 $\pm$ .001
NBDep	49	-223366.7 $\pm$ 4.9	.401 $\pm$ .005	1.63 $\pm$ 0.07	.019 $\pm$ .001

- GenRev: General reversible, neighbour-independent
- GenIrr: General irreversible, neighbour-independent
- GenIrr+CG: Dependence through $CG \rightarrow TG/CA$ substitution only.
- GenIrr+CG+Dinuc: Dependence through $CG \rightarrow TG/CA$, single dinucleotide rate.
- NBDep: General neighbour dependent model, single dinucleotide rate.

Simultaneous double-nucleotide substitutions?

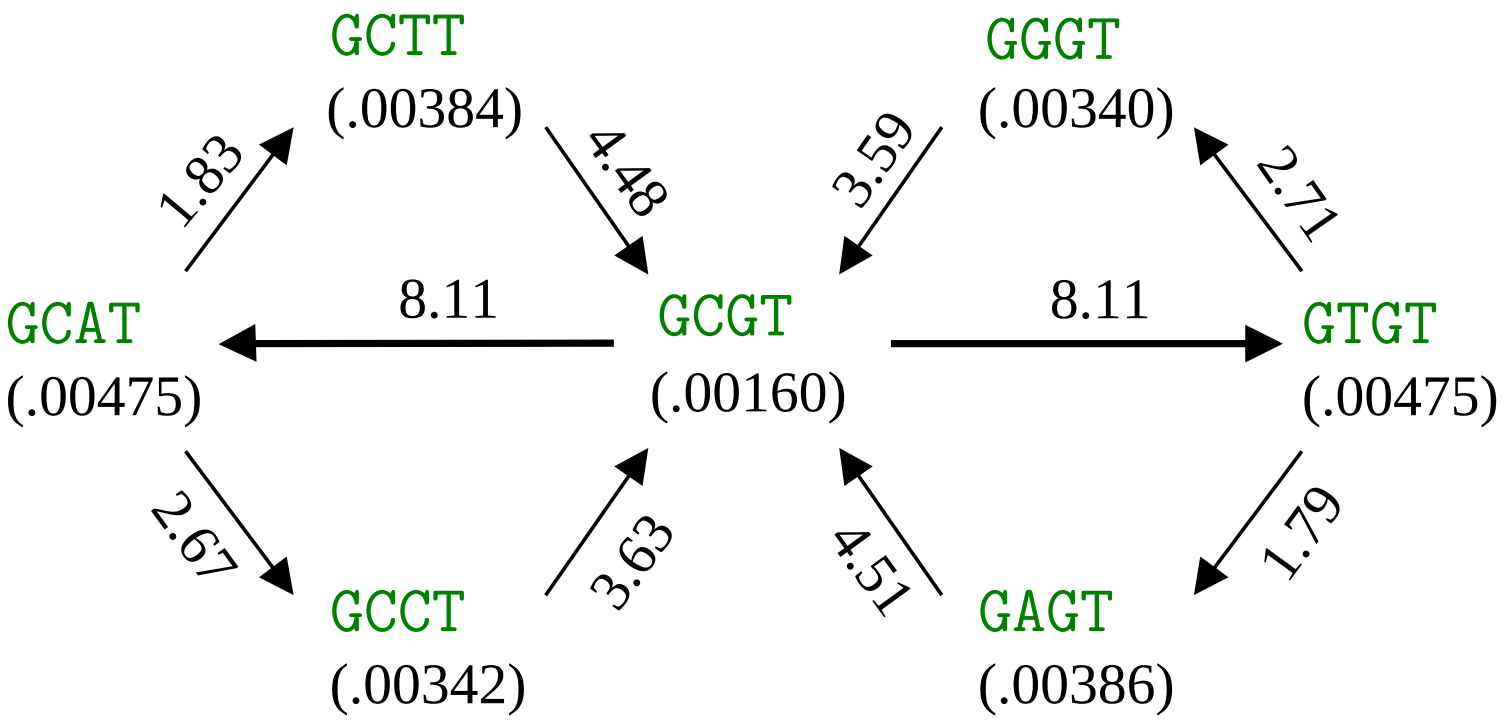
Averof, Rokas, Wolfe and Sharp, *Evidence for a High Frequency of Simultaneous Double-Nucleotide Substitutions*, Science 2000.

- **Closely related DNA:** Multiple alignment of 7 primates
  Observed: 732 substitutions, 30 tandems, expected ≈ 15.
- **Divergent DNA:** Codon switches in highly conserved serines (`TCN`, `AGY`) among range of eukaryotes and prokaryotes.

Both methods: doublet rate ≈ 0.1 per site per Gyr $= 2\%$ of singlet rate.

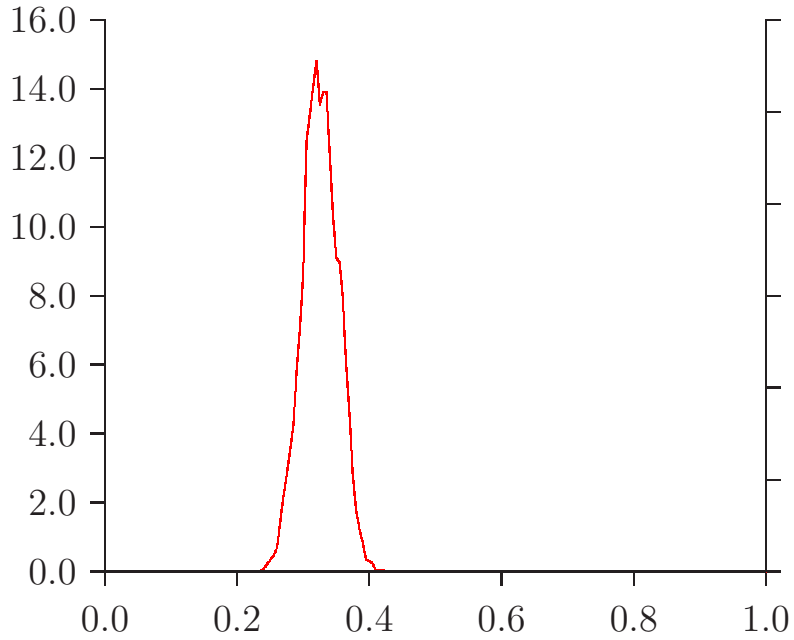
Corresponds to ≈ 0.015 per site per 150 Myr; cf. our estimate: ≈ 0.02.

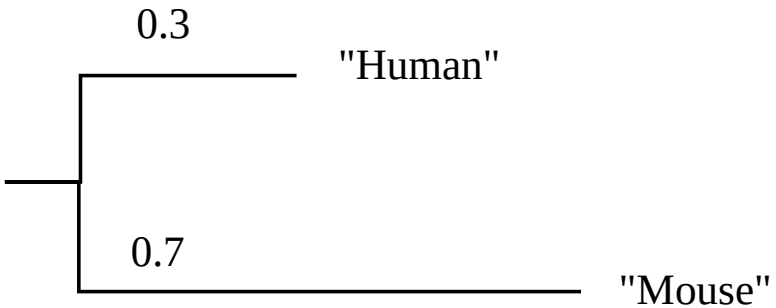
Irreversibility



- (.00160) etc.: Equilibrium probability of sequence
- 8.11 etc.: Net equilibrium flow $\times 10^{-4}$

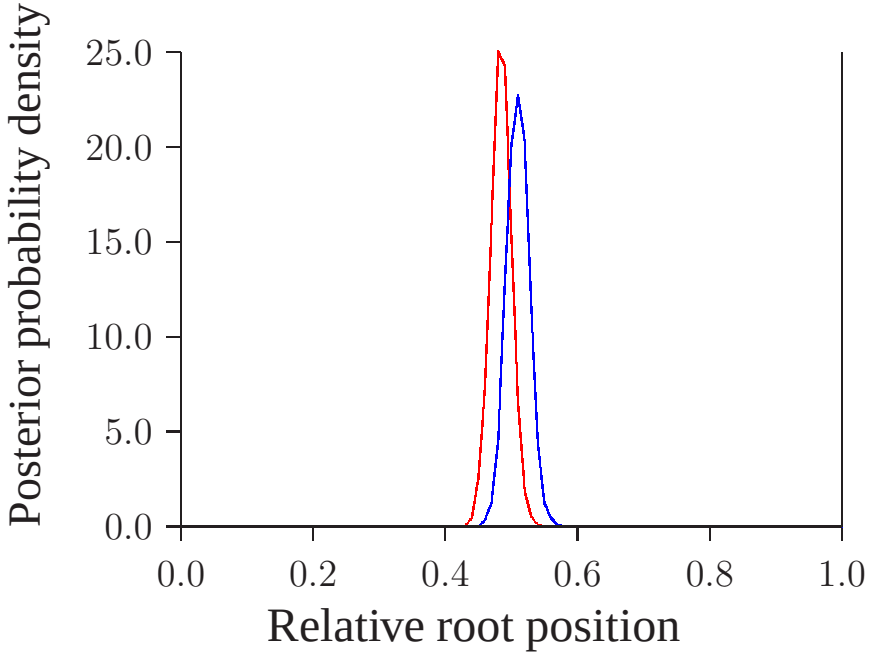
Root inference - synthetic data

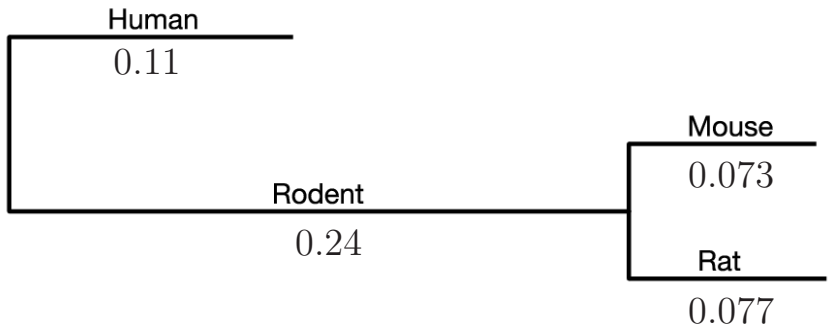




Root position inference 0.33 ± 0.03 (true value 0.3)

Root inference





Tree estimate: Nature April 2004, Rat Genome Sequencing Project Consortium

Inferred root positions:

0.484 ± 0.014 (chromosome 21), 0.510 ± 0.016 (chromosome 10).

Discussion

- Approximation method sufficiently accurate for reliable rate estimates in mammalian phylogeny.

Discussion

- Approximation method sufficiently accurate for reliable rate estimates in mammalian phylogeny.

- CpG effect accounts for strongest context effect.
  Many more significant context effects exist.

Discussion

- Approximation method sufficiently accurate for reliable rate estimates in mammalian phylogeny.

- CpG effect accounts for strongest context effect.
  Many more significant context effects exist.

- Unexpected "molecular clock" rooting of human-mouse tree:
  - Underlying hypothesis very likely not true.
    (Substitution process on the two lineages identical except for scale change.)
  - Results suggest that "irreversible component" may be clock-like.